Sens, synonymes et définitions

Ingrid Falk[1] Claire Gardent[2] Evelyne Jacquey[3] Fabienne Venant[4]
(1) INRIA / Université Nancy 2
(2) CNRS / LORIA, Nancy
(3) CNRS / ATILF, Nancy
(4) Université Nancy 2
ingrid.falk@loria.fr,claire.gardent@loria.fr, evelyne.jacquey@atilf.fr,
fabienne.venant@loria.fr

Résumé. Cet article décrit une méthodologie visant la réalisation d'une ressource sémantique en français centrée sur la synonymie. De manière complémentaire aux travaux existants, la méthode proposée n'a pas seulement pour objectif d'établir des liens de synonymie entre lexèmes, mais également, d'apparier les sens possibles d'un lexème avec les ensembles de synonymes appropriés. En pratique, les sens possibles des lexèmes proviennent des définitions du TLFi et les synonymes de cinq dictionnaires accessibles à l'ATILF. Pour évaluer la méthode d'appariemment entre sens d'un lexème et ensemble de synonymes, une ressource de référence a été réalisée pour 27 verbes du français par quatre lexicographes qui ont spécifié manuellement l'association entre verbe, sens (definition TLFi) et ensemble de synonymes. Relativement à ce standard étalon, la méthode d'appariemment affiche une F-mesure de 0.706 lorsque l'ensemble des paramètres est pris en compte, notamment la distinction pronominal / non-pronominal pour les verbes du français.

Abstract. We present a method for grouping the synonyms of a word into sets representing the possible meanings of that word. The possible meanings are given by the definitions of a general dictionary for French, the TLFi (Trésor de la langue française informatisé) and the method is applied to the synonym of 5 synonym dictionnaries. To evaluate the method, we manually constructed a gold standard where for each (word, definition) pair, 4 lexicographers specified the set of synonyms they judge adequate. The method scores an F-measure of 0.602 when no distinction is made between pronominal and non-pronominal use and 0.706 when it is.

1 Introduction

L'importance de la sémantique lexicale pour le TAL n'est plus à démontrer. La question de la représentation du sens et de l'accès à l'information dans les ressources lexicales devient fondamentale pour la résolution de nombreux problèmes actuels tels que la construction d'ontologies, la recherche d'information et la génération de textes. Cette question est d'autant plus importante que pour le français, on manque encore de ressources sémantiques de qualité, structurées, librement accessibles et non limitées à un domaine spécialisé.

Dans cet article, nous présentons une méthode visant à regrouper les synonymes d'un mot en sous-ensembles correspondant aux différents sens possibles de ce mot. Les sens possibles sont donnés par les définitions du TLFi et l'objectif est d'appliquer cette méthode à la base des synonymes de l'ATILF. Par exemple, étant donné les définitions TLFI (2, en italiques) et les synonymes (1) extraits des dictionnaires de la base de l'ATILF pour le verbe *achever*, la méthode proposée vise à associer aux définitions TLFI, le sous-ensemble de synonymes approprié (2).

(1) *Achever :* abattre, aboutir, accomplir, aiguiser, améliorer, anéantir, assommer, boucler, cesser, clore, clôturer, complémenter, compléter, conclure, conduire, consommer, continuer, couronner, estoquer, expédier, exécuter, finir, parachever, parfaire, perfectionner, raser, ruiner, réaliser, réussir, se taire, terminer, tuer, être

(2) a. *Mettre la dernière main pour perfectionner :* accomplir, boucler, cesser, clore, clôturer, conclure, perfectionner, raser

b. *Porter un coup mortel à un animal déjà atteint physiquement ; donner le coup de grâce :* assommer, couronner, estoquer

c. *Porter un coup mortel à un animal déjà atteint physiquement ; donner le coup de grâce :* assommer, couronner, estoquer

d. *Mener à sa fin, compléter l'action de :* accomplir, cesser, complémenter, compléter, conduire, finir, parfaire

Pour évaluer notre méthode, nous avons créé manuellement une ressource de référence. Relativement à cette ressource, notre méthode affiche une F-mesure de 0.602 lorsque la distinction prononominal/non-pronominal n'est pas prise en compte, et de 0.706 lorsqu'elle l'est.

L'approche proposée présente des similarités avec d'une part, le travail d'extraction d'un wordnet libre pour le Français (WOLF) présenté dans (Sagot & Fiser, 2008) et d'autre part, l'élaboration de DicoSyn[1] (Manguin *et al.*, 2004). Elle diffère de DicoSyn en ce qu'elle permet d'associer chaque groupe de synonymes avec une définition représentant le sens de ces synonymes. Elle diffère par ailleurs de l'approche présentée dans (Sagot & Fiser, 2008) en ce que les ensembles de synonymes (« synsets ») obtenus sont créés à partir de données francophones (définitions TLFI et synonymes) plutôt que par traduction à partir du WordNet de Princeton pour l'anglais et désambiguisation sur des données multilingues.

L'article est structuré de la façon suivante. La section (2) présente la méthode mise au point pour regrouper sens et synonymes, la section (3) résume les résultats obtenus, la section (4) compare l'approche proposée avec les approches adoptées pour la création de DicoSyn et du WOLF et la section (5) indique les perspectives de recherche pour l'avenir.

2 Méthode

Pour un même lexème, les différents dictionnaires de synonymes de l'ATILF donnent généralement un nombre de sens et des ensembles de synonymes distincts. Afin de traiter de ces divergences et de fusionner l'information contenue par chacun de ces dictionnaires, nous prenons les définitions du TLFI comme répertoire des sens possibles d'un lexème et nous utilisons

[1] http://www.crisco.unicaen.fr/Presentation-du-dictionnaire.html

des mesures de similarité pour associer chacun des synonymes répertoriés dans les différents dictionnaires de synonymes avec les sens (définitions du TLFi) adéquats. La méthode employée comprend les trois grandes étapes suivantes :
- *Extraction des définitions du* TLFI *et création d'un index par définition.*
- *Mesure de la similarité entre index.* Etant donné un verbe V, un synonyme S_V^i de V, les index extraits des définitions TLFI $D_V^1, ..., D_V^n$ de V et l'union des index extraits des définitions D_{S_V} de V, des mesures de similarité sont appliquées pour déterminer la définition D_V^j de V la plus similaire des définitions de S_V^i.
- *Association Synonyme/Définition.* Chaque synonyme est associé avec la (les) définition(s) D_V^i de V la plus similaire à l'ensemble de ses définitions.

2.1 Traitement des définitions et création des index

Parce que les calculs de similarité opérent sur les index créés à partir des définitions TLFI, il est important de normaliser et d'enrichir ces définitions autant que possibl. Il importe par exemple, d'éviter que deux formes soient considérées comme distinctes alors qu'elles représentent le même lemme (*chat/chats*) ou encore de pallier le fait que certaines définitions du TLFI sont vides de contenu parce que le texte de la définition est supposé donné par héritage à partir des niveaux supérieurs de l'entrée dictionnairique. Nous détaillons dans ce qui suit, la structure des entrées du TLFI et indiquons comment cette structure est utilisée pour construire les index sur lesquels seront appliquées les mesures de similarité.

Les différents types d'information d'une entrée du TLFi L'information fournie dans le TLFi se subdivise en trois grands ensembles : (1) différents types d'informations non rétroconverties au format XML donc non exploitables ici (étymologie, fréquences dans le corpus de référence du dictionnaire, dérivés, orthographe et prononciation) ; (2) l'information relative à l'organisation hiérarchique (marques de niveau hiérarchique, marques de plan, retraits, puces et tirets) ; et (3) l'information lexicographique proprement dite qui est divisée en blocs d'information cohérents pouvant contenir : un *domaine d'emploi* (`ferronnerie, justice`, etc.), un *indicateur d'emploi* (`figuré, familier, pronominal` etc), des *crochets* (informations très diverses), une *définition* précédée ou non d'une *locution définie* et/ou une relation de *synonymie ou antomymie*. C'est ce dernier type d'information (l'information lexicographique) qui fournit l'essentiel des informations utilisées pour construire les index. L'information hiérarchique, est pour sa part, essentiellement exploitée pour mettre en œuvre l'héritage descendant défini ci-dessous.

Identification et catégorisation des blocs définitionels Afin d'identifier et de catégoriser les définitions d'une entrée du TLFi, les règles d'extraction suivantes ont été appliquées :

Détermination du nombre d'emplois ou sens de chaque lexème : est considéré comme un emploi général distinct, tout bloc d'information contenant une définition, et/ou une indication de domaine technique, et/ou un synonyme ou antonyme ;

Différenciation des emplois pronominaux et non-pronominaux : est considéré comme un emploi pronominal, tout bloc d'information contenant un indicateur d'emploi contenant l'information *pronominal* si le code grammatical de l'entrée du verbe est ambigu, ou bien, tous les blocs d'information de l'entrée d'un verbe indiqué comme pronominal dans son code grammatical ;

Identification des emplois figés : est considéré comme la description d'un emploi figé ou semi-figé, tout bloc d'information dans lequel une expression ou locution définie précède une définition.

Héritage de l'information. Nous avons procédé à un héritage descendant et contrôlé afin d'expliciter l'information laissée implicite par le lexicographe. Parmi les différents types d'informations, certains ont été considérés comme strictement locaux, d'autres comme pouvant être propagés sur les blocs hiérarchiquement dominés.

Informations héritables : les indicateurs d'emploi (par exemple *emploi pronominal*), les crochets, les indications de domaine technique, le code grammatical de l'entrée ;

Informations locales : les définitions, les synonymes et antonymes, les expressions illustratives et les exemples.

Création des index. Pour chaque définition identifiée à partir des informations hiérarchiques, l'information extraite du TLFi comprend : le texte de la définition, les synonymes et les antonymes éventuellement associés à cette définition et par héritage, les indicateurs de domaine technique. A partir de cette information, l'index créé est la séquence des lemmes[2] de catégorie : Inf (infinitif), S (substantif), A (adjectif, sauf cas cardinal, part. passé, adverbe), APs (adverbes), Pr (participe présent sauf gérondif), Ps (part. passé sauf cas infitinitif). Par exemple, dans le cas des définitions du verbe *projeter*, les index extraits seront les suivants : la définition *Jeter loin en avant avec force* est associée à l'index ⟨ jeter, loin, avant, force ⟩, la définition *CIN. AUDIOVISUEL. Passer dans un projecteur*, qui contient un domaine, à l'index ⟨ cinéma, audiovisuel . passer, projecteur ⟩, et la définition *Eclaircir. Synon. jeter quelque lumière* qui comporte un synonyme, à l'index ⟨ éclaircir. jeter, lumière ⟩.

2.2 Calculs des regroupements synonymiques

Pour chaque synonyme SYN_V d'un verbe V nous cherchons à déterminer la définition TLFi de V qui est la plus similaire des définitions de SYN_V.[3] Pour ce faire, nous appliquons des mesures de similarité entre l'index extrait de chaque définition de V et l'index représentant l'ensemble des définitions de SYN_V. Le synonyme SYN_V sera associé aux définitions D_V^i de V pour lesquelles la mesure de similarité est la meilleure. Afin d'estimer l'impact de la mesure de similarité utilisée, nous avons comparé les 6 mesures de similarité suivantes :

1. *Recouvrement de mots* (Simple word overlap) : le score obtenu est le nombre de mots communs aux deux textes comparés.
2. *Recouvrement de mots étendu* (Extended word overlap). À une séquence de n mots commune entre les textes comparés est attribuée un score de n^2. Le score donné par la mesure de recouvrement de mots étendue (Banerjee & Pedersen, 2003)[4] est la somme des scores de recouvrements de n mots.
3. *Recouvrement de mots étendu avec normalisation* (Extended word overlap, normalised). Le score obtenu par le *Recouvrement de mots étendu* (2) est normalisé par le nombre de mots des définitions comparées.

[2]Les définitions TLFi sont déja lemmatisées et catégorisées et les lemmes sont aussi accessibles dans le XML.

[3]Le sens du synonyme considéré n'étant pas identifié par les dictionnaires de synonymes, il est impossible de déterminer la ou les définitions pertinentes pour la comparaison verbe/synonyme. C'est pourquoi prenons comme base de comparaison l'ensemble des définitions du synonyme et non, la ou les définitions correspondant au sens impliqué. Bien que cette procédure soit indéniablement inexacte, les résultats obtenus indiquent que le bruit introduit reste acceptable.

[4](Banerjee & Pedersen, 2003) inclut en outre dans la mesure de similarité les définitions des hyponymes, hyperonymes, méronymes, holonymes et troponymes donnés par le Princeton WordNet pour les mots impliqués dans les textes comparés. Nous n'avons pas pu prendre en compte cette extension du fait de la couverture relativement pauvre de l'EuroWordNet Français (2349 verbes contre e.g., 5077 verbes pour Le Petit Robert). Le WOLF (Sagot & Fiser, 2008) contient 979 verbes (1544 synsets verbaux) et est imparfait car résultant d'un processus d'acquisition automatique. Il serait néanmoins intéressant de voir dans quelle mesure l'information qu'il contient impacte la qualité des résultats.

4. *Vecteurs du premier ordre*. Les dimensions de l'espace des mots sont les mots pleins des définitions des verbes dans le TLFI. Les coefficients des vecteurs v_{D_V} et v_{S_V} associés sont pondérés sur un schéma de type tf-idf [5]. Le score de similarité est le produit scalaire de v_{D_V} et v_{S_V}.
5. *Vecteurs du second ordre sans seuil tf.idf*. Les dimensions de l'espace des mots sont tous les mots pleins de toutes les définitions de tous les verbes du TLFI. Le vecteur v_T associé à un texte T est la somme (vectorielle) des *Vecteurs du premier ordre* (4) pour chaque mot de T. Le score de similarité est le produit scalaire des vecteurs résultants.
6. *Vecteurs du second ordre avec seuil tf.idf* Comme (5), mais pour construire l'espace de mots, seules les mots pleins avec un tf-idf supérieur à 1500 sont pris en considération.

Etant donné une mesure de similarité $\mathcal{M}$ et l'ensemble $\text{DEF}(V)$ des définitions TLFi de V, on calcule les scores de similarité $Sim_{\mathcal{M}}(\text{IDX}(S_V), I_V^i)$ pour chaque index I_V^i.

On considère que S_V est un synonyme de V pour le sens représenté par la définition D_V^i si

$$\begin{array}{rcl} Sim_{\mathcal{M}}(\text{IDX}(S_V), I_V^i) & > & 0 \quad \text{et} \\ Sim_{\mathcal{M}}(\text{IDX}(S_V), I_V^i) & = & \max\limits_{D_V^i \in \text{DEF}(V)} Sim_{\mathcal{M}}(\text{IDX}(S_V), I_V^i) \end{array}$$

Autrement dit, on associe le synonyme aux définitions pour lesquelles on obtient le meilleur score de similarité pour autant qu'il ne soit pas nul. La Figure 1 illustre les résultats de cette méthode pour le verbe *abandonner*[6]. 1(a) montre les scores de similarité de 3 synonymes avec la définition TLFI *Quitter un lieu, ne plus l'occuper.* du verbe *abandonner*, calculés par 3 mesures de similarité différentes. Dans le tableau 1(b), la mesure de similarité est fixe (*Recouvrement de mots étendu avec normalisation*) mais les scores sont affichés pour un synonyme et plusieurs définitions. On voit par exemple que le synonyme *déloger* atteint le score maximal pour la définition 1.1.1.1.1.2.1 (*Quitter un lieu* ...). Il sera considéré synonyme à *abandonner* avec le sens représenté par cette définition. Enfin 1(c) montre les ensembles de synonymes obtenus par la méthode exemplifiée ci-dessus. Par rapport à la référence, de l'ensemble des 6 synonymes attribués au sens 1.1.1.1.2.1 (*Renoncer* ...) par le système, 5 ont aussi été attribués à ce sens par un lexicographe[7].

3 Résultats

Afin d'évaluer la qualité des résultats obtenus, de comparer les mesures de similarité et de mesurer l'impact du pré-traitement linguistique, nous avons manuellement créé un échantillon de référence permettant de calculer le rappel et la précision de la méthode proposée et de différentes variantes de cette méthode (différentes mesures de similarité, différents degré de pré-traitement linguistique).

L'échantillon de référence. Nous avons sélectionné un échantillon de verbes variant en terme de fréquence, polysémie et généricité. Pour chaque dimension, nous avons identifié des verbes illustrant trois valeurs (basse, moyenne et élevé) induisant un total de 27 verbes (3^3). La fréquence est estimée à partir d'une liste de fréquence extraite d'une analyse avec Syntex de 10

[5] Le coefficient v_i, correspondant à la dimension du mot m_i est calculé selon le schéma suivant :

$$v_i = \frac{\text{nbr. d'occ. du mot } m_i \text{ dans } D_V}{\text{nbr. de définitions dans l'espace de mots comprenant } m_i}$$

[6] Par manque de place, nous n'avons inclus ni toutes les définitions, ni tous les synonymes

[7] Pour les deux autres exemples présentés ici l'accord lexicographes/système est de $5/6$ et $5/7$ respectivement

Synonyme	r	re	ren
déguerpir	1	1	0.067
déloger	5	8	0.778
déménager	2	5	0.722

(a) Scores de similarité selon les mesures *recouvrement de mots* (r), *recouvrement étendu* (re) et *recouvrement étendu avec normalisation* (ren) de quelques synonymes par rapport à la définition *Quitter un lieu, ne plus l'occuper*.

Synonyme	1.1.1	1.1.1.2.1	1.1.1.1.2.1	1.1.1.1.1.2.1
rompre	**0.213**	0.060	0.000	0.000
déloger	0.000	0.042	0.000	**0.778**
se désister	0.000	0.000	**0.167**	0.000

(b) Scores de similarité par synonyme et définition pour quelques synonymes par rapport à quelques définitions, données par leurs identifiants. La mesure de similarité est *recouvrement de mots étendu avec normalisation*. En gras les valeurs maximales par synonyme et l'ensemble de définitions.

1.1.1 : *Rompre le lien qui attachait à une chose ou à une personne.*   *abjurer*, accorder, *céder, délaisser, déménager, se désister, exposer*, **finir**, fuir, *larguer, livrer, lâcher, négliger*, planter, *rejeter, renier, résigner*, rompre, sacrifier, *semer, tomber*
1.1.1.1.2.1 : *Renoncer à poursuivre une action, une recherche ; renoncer à une entr., à un projet.*   confier, se démettre, **démissionner**, se départir, se désister, *dételer, flancher, laisser, lâcher, plaquer*, reculer, *renoncer, se rétirer*
1.1.1.1.1.2.1 : *Quitter un lieu, ne plus l'occuper ;*   *donner, déguerpir*, **déloger**, **déménager**, **déserter**, *laisser*, lâcher, *quitter*, rabattre, se séparer, **évacuer**

(c) Groupements de synonymes par définition. La mesure de similarité est *recouvrement de mots étendu avec normalisation*.

FIG. *1:* Mesures et scores de similarité pour quelques synonymes de *abandonner*. Exemple de calculs de regroupement de synonymes.

ans du Monde (Bourigault & Fabre, 2000), la généricité à partir de la position du verbe dans le module français de l'EuroWordNet et la polysémie à partir du nombre de définitions listées dans le TLFI. Pour chacun des verbes sélectionnés, chaque définition TLFi a été associée manuellement à un ou plusieurs synonymes de la base par quatre lexicographes ou linguistes. Sur un total de 7 047 triplets ⟨ Verbe, Définition, Synonyme ⟩, le pourcentage d'accord inter-annotateurs varie entre 74 et 87% pour les paires d'annotateur et descend à 63% pour les 4 annotateurs. Aucune des paires ne présente un accord parfait ce qui montre la difficulté de la tâche. Pour l'évaluation, nous utilisons la référence créée par le premier annotateur de la paire ayant l'accord le plus élevé (87%).

Variantes et résultats. Dans un premier temps, l'échantillon de référence a été utilisé pour évaluer l'impact de la mesure de similarité utilisée sur le rappel et la précision. Le rappel est le nombre de triplets $\langle$ Verbe, Def_i, SYN_{Verbe} $\rangle$ trouvés par le système et contenus dans la référence divisé par le nombre de triplets contenus dans la référence. La précision est le nombre de triplets $\langle$ Verbe, Def_i, SYN_{Verbe} $\rangle$ trouvés par le système et contenus dans la référence divisé par le nombre de triplets trouvés par le système. Les résultats obtenus pour les différentes mesures sont donnés en Tableau 1, avec pour ligne de base (baseline), une association arbitraire entre synonymes et paires $\langle$ Verbe, Def_i $\rangle$.
Dans la première série de résultats (sans prise en compte de la distinction pronominal vs. non-pronominal), on observe que les résultats obtenus par les mesures de similarité sont meilleurs que le baseline. Les mesures utilisées permettent donc d'avoir des résultats meilleurs que le hasard. On observe également qu'il y a peu de différence entre les différentes mesures et que la précision reste relativement basse.
Cependant, dans la mesure où l'extraction des informations lexicographiques du TLFI a pris en

compte la distinction pronominal vs. non-pronominal, il a été possible de séparer les emplois pronominaux des emplois non pronominaux. Les synonymes de la forme pronominale ne sont alors comparés qu'aux emplois pronominaux et vice versa. Les résultats obtenus sur la base de cette nouvelle procédure sont donnés dans la seconde série de résultats, (Tableau 1). Ils montrent un gain net en précision comme en rappel avec une F-mesure de 0.706 (pour le recouvrement étendu normalisé) contre 0.602 sans la distinction pronominal/non pronominal.

	Sans diff. pron vs. non-pron			**Avec diff. pron vs. non-pron**		
Mesure sim.	**R**	**P**	**F**	**R**	**P**	**F**
baseline	0.497	0.315	0.385	0.44	0.433	0.437
Recouvrement	0.725	0.508	0.598	0.697	0.685	0.691
Rec. étendu	0.723	0.508	0.597	0.697	0.685	0.691
Rec. étendu norm.	**0.729**	**0.513**	**0.602**	**0.711**	0.67	**0.706**
Vecteurs 1er ordre	0.727	0.51	0.56	0.704	0.693	0.698
Vecteurs 2nd ordre, sans tf.idf	0.715	0.503	0.59	0.698	0.686	0.692
Vecteurs 2nd ordre, avec tf.idf	0.717	0.505	0.592	0.701	**0.689**	0.695

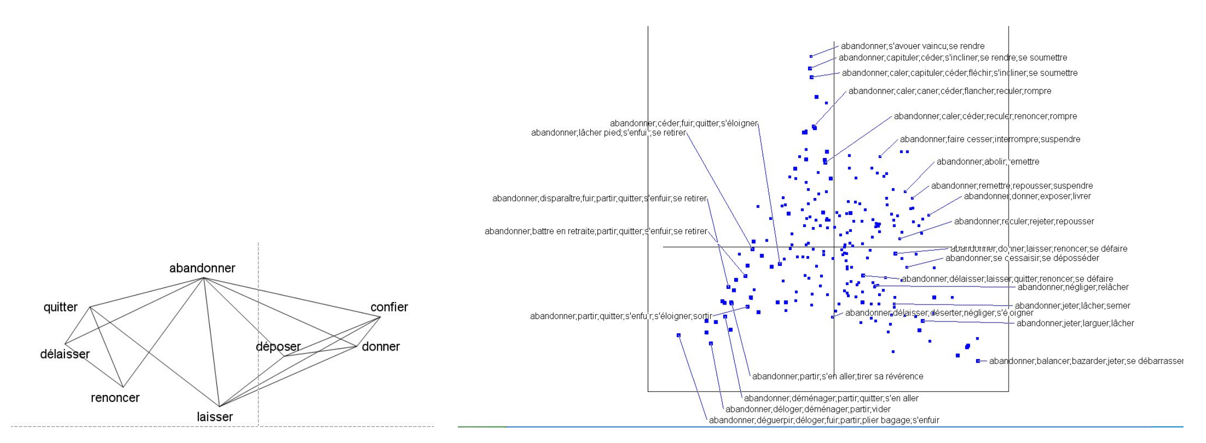
TAB. *1:* Précision, rappel et F-mesure pour les différentes mesures de similarité, avec et sans la distinction pronominal / non-pronominal.

En résumé, la méthode proposée permet une nette amélioration par rapport à un assignement synonymes/définitions fait au hasard (F-Mesure de 0.706% vs. 0.437%). La différence entre résultats avec et résultats sans prétraitement des pronominaux montrent en outre qu'un prétraitement fin des définitions du TLFI est une composante importante pour des résultats de qualité.

4 Comparaison / interaction possibles avec les ressources françaises existantes

Dicosyn. Dicosyn est un dictionnaire électronique, disponible en ligne. Il est issu de la fusion de 7 dictionnaires[8].Le travail initial a été réalisé par (Ploux & Victorri, 1998). On explore ce dictionnaire sous la forme d'un graphe : les sommets sont les entrées du dictionnaire. Un lien est tracé entre deux sommets lorsque le dictionnaire atteste de leur synonymie. Pour étudier la sémantique d'un mot donné, on extrait du graphe de synonymie le sous-graphe constitué par ce mot et tous ses synonymes, et seulement ses synonymes. La construction des espaces sémantiques repose sur le même constat que celui évoqué précédemment : un synonyme ne suffit pas en général pour définir un sens du mot étudié. La Figure 2(a) montre un extrait du graphe de synonymie de *abandonner*. On y voit que *laisser*, est à la fois synonyme de *quitter* et de *confier*, qui correspondent à deux sens différents de *abandonner*. L'idée est donc, ici aussi, de caractériser un sens par un ensemble de synonymes. Plus précisément, Dicosyn utilise les cliques du graphe de synonymie. Une clique est un ensemble de sommets deux à deux synonymes le plus grand possible. Le graphe de la Figure 2(a) présente ainsi 2 cliques : ⟨ *abandonner ; confier ; donner ; déposer, laisser* ⟩, ⟨ *abandonner ; délaisser ; laisser ; quitter ; renoncer ;* ⟩ L'idée sous-jacente à la construction des espaces sémantiques est qu'une clique correspond, en première approximation, à une nuance de sens possible pour le mot considéré. Nous ne détaillons pas ici la technique de construction de l'espace sémantique cf. (Ploux & Victorri, 1998). Disons simplement que l'espace sémantique est l'espace euclidien engendré par *abandonner* et tous ses synonymes. Les points de cet espace sont les cliques du graphe. Les coordonnées d'une clique dépendent des synonymes qu'elle contient. La Figure 2(b) montre

[8]http ://www.crisco.unicaen.fr/cgi-bin/cherches.cgi



(a) Un extrait du graphe de synonymie de *abandonner*.

(b) Espace sémantique associé au verbe *abandonner*

FIG. 2: *abandonner* : un extrait du graph de synonymie (2(a)), l'espace sémantique associé (2(b)).

l'espace sémantique associé à abandonner. Il s'agit d'une vision partielle, puisque l'espace étant de grande dimension, il faut le projeter pour pouvoir en obtenir une vision 2D. Nous voyons ici la projection selon les deux premiers axes de l'analyse en composantes principales. De telles visualisations permettent une visualisation globale de la sémantique d'une unité. Leur intérêt principal est de rendre compte de la continuité du sens, de la difficulté à tracer des frontières nettes entre les sens, de la « ressemblance de famille » existant entre les différents sens d'une unité. Sur la Figure 2 on voit ainsi se dégager 4 pôles de sens, et la façon dont on peut passer continûment d'un de ces pôles à un autre, en passant par des nuances de sens intermédiaires. Comparons ces pôles aux 7 sens principaux que l'on peut extraire du TLFI :

1. Rompre le lien qui attachait à une chose ou une personne ; renoncer à un pouvoir, à des doirs, à la possession d'un bien ou à l'utilisation d'une chose
2. Quitter un lieu ne plus l'occuper : abandonner un pays, un domicile
3. Cesser de défendre une cause, renoncer à des principes, à une idée en la rejetant ou simplement en s'en séparant : abandonner une cause, des principes
4. Renoncer à poursuivre une action, une recherche, renoncer à une entreprise, un projet : abandonner des recherches
5. Quitter qqn, s'en séparer ; laisser qqn à lui-même, le laisser seul
6. Laisser à qqn la possession ou le soin d'un bien (ou d'une pers.), laisser qqc. à l'entière disposition de qqn : abandonner à qqn le soin de faire qqc
7. Laisser qqc. ou qqn en proie à qqc (gen. une force hostile) : abandonner une terre aux ravages, abandonner un enfant au bagne

Toute la partie supérieure de l'espace sémantique rassemble des synonymes comme *s'avouer vaincu*, *se rendre*, *caler*, *céder*, *interrompre*, *repousser* qui correspondent au sens 4. En descendant vers la droite de l'espace sémantique, on arrive, via la notion de renoncement et de dépossession, à des synonymes comme *jeter*, *lâcher*, *larguer*, *jeter* qui correspondent (selon que l'objet est humain ou non) aux sens 1 et 5. La partie inférieure gauche de l'espace sémantique rassemble des synonymes comme *déguerpir*, *fuir* qui correspondent plutôt au sens 2. Si une telle exploration continue peut faire le bonheur d'un lexicographe ou d'un sémanticien, on voit qu'elle est plus difficile à exploiter automatiquement de par sa granularité trop fine. Les expériences de désambiguïsation menées à partir des espaces sémantiques de dicosyn (Venant, 2007; Venant, 2008) ont nécessité un regroupement manuel des cliques, les nuances de sens, en sens macroscopiques correspondant à des définitions du TLFI. De plus, comme on a pu le voir ici, on n'a pas accès à des discriminations de sens en fonction de la nature de l'objet ou du sujet. On voit donc ici toute la complémentarité des deux approches. L'idéal serait à terme d'établir

des correspondances entre les deux ressources, en projetant sur les cliques les appariements synonymes/définitions que nous avons obtenus.

WOLF. Le WOLF (WOrdnet Libre du Français est une ressource lexicale sémantique pour le français, librement disponible (Sagot & Fiser, 2008). Il s'agit d'un wordnet construit à partir du Princeton WordNet (PWN) et de diverses ressources multilingues (Sagot & Fiser, 2008), au moyen de deux approches complémentaires. Comme dans tout WordNet, les unités lexicales sont réparties en catégories et organisées en une hiérarchie de nœud, structurée par des relations sémantiques. Chaque nœud représente un concept et est associé à un synset (ensemble de synonymes dénotant ce concept). Le WOLF reprend la structure de PWN. Pour chaque lexème monosémique du PWN, on crée l'entrée française correspondante, via une ressource bilingue anglais-français. Les lexèmes polysémiques ont été traités au moyen d'une approche reposant sur l'alignement en mots d'un corpus parallèle en cinq langues, dont le français. Différents lexiques multilingues ont été extraits de ce corpus aligné. La désambiguïsation sémantique des ces lexiques s'appuie sur le fait que les Balkanet wordnet des langues concernées utilisent les mêmes identifiants de synset.
Le Wolf ainsi créé contient tous les synsets du PWN, y compris ceux pour lesquels aucun lexème français correspondant n'a été trouvé. Le travail que nous avons réalisé est complémentaire de l'approche utilisée par Sagot et al. La méthode qu'ils ont choisie permet de créer automatiquement un WordNet à partir d'un autre existant, mais elle repose sur l'hypothèse d'une structuration sémantique identique en anglais et en français. Nous exploitons quant à nous la structure sémantique du TLFI. Cette structure présente l'avantage d'être spécifique au français. Le point commun entre les deux approches réside dans l'association de chaque concept avec un ensemble de synonymes. On doit pouvoir envisager une interaction entre les deux ressources. Par exemple, une comparaison des deux structures permettrait une enrichissement mutuel : d'une part la validation, l'enrichissement ou la correction des synsets existants, d'autre part l'exploitation des relations sémantiques tirées du WordNet. Ces informations sont sans doute présentes dans le TLFI, mais elles sont très difficiles à extraire et exploiter automatiquement.

5 Conclusion

Nous avons présenté une méthode permettant d'associer synonymes et définitions. Afin d'évaluer la qualité des résultats obtenus, un échantillon de référence a été créé manuellement avec un accord inter annotateurs variant entre 63% et 87%. Les résultats obtenus pour les différentes variantes explorées (différentes mesures de similarité, différents niveaux de prétraitement des définitions) montrent d'une part, que la méthode permet une précision proche de la borne inférieure délimitée par l'accord inter-annotateur et d'autre part, qu'un prétraitement poussé des définitions et en particulier, la différentiation entre usage pronominal et non pronominal, est essentiel pour la bonne qualité des résultats.
Nous comptons améliorer la méthode proposée par un pré-traitement plus poussé des définitions du TLFI. En particulier, la prise en compte des informations de sous-catégorisation devrait permettre à la fois d'enrichir les index utilisés par les calculs de similarité et de faire le lien avec les lexiques syntaxiques répertoriant les schémas valenciels des éléments prédicatifs tels que Lefff (Sagot *et al.*, 2006), Synlex (Gardent *et al.*, 2006), Treelex[9] (Kupsc & Abeillé, 2008), LexSchem[10] (Messiant *et al.*, 2008) et Dicovalence[11] (Eynde & Mertens, 2003).

[9] http://erssab.u-bordeaux3.fr/article.php3?id_article=150
[10] http://www-lipn.univ-paris13.fr/~messiant/lexschem.html
[11] http://bach.arts.kuleuven.be/dicovalence/

Ensuite, nous pensons appliquer la méthode aux dictionnaires synonymiques disponibles dont en particulier, les dictionnaires des synonymes de l'ATILF et le wiktionnaire du français.
Enfin, nous souhaitons mettre au point une méthodologie permettant de tirer parti des complémentarités de WOLF et de DicoSyn pour évaluer et complémenter le dictionnaire synonymique résultant.

Remerciements. Nous remercions Christiane Jadelot, Aurélie Merlot et Mick Grzesitchak pour leur participation à la création de l'échantillon de référence, Rohan Railkar pour sa contribution à l'implantation du processus d'héritage et l'ATILF pour la mise à disposition des sources dictionnairiques (TLFi et Base des synonymes). Ce projet a été partiellement financé par le pôle TALC (Traitement automatique des langues et des connaissances,talc.loria.fr) du contrat plan Etat Région MISN (Modélisation, information et systèmes numériques).

Références

BANERJEE S. & PEDERSEN T. (2003). Extended gloss overlaps as a measure of semantic relatedness.

BOURIGAULT D. & FABRE C. (2000). *Approche linguistique pour l'analyse syntaxique de corpus*. Rapport interne, Université Toulouse - Le Mirail. Cahiers de Grammaires, no. 25.

EYNDE K. V. D. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, **13**, 63–104.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2006). Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Traitement Automatique de la Langue Naturelle - TALN 2006 Actes de la 13ème conférence sur le Traitement Automatique de la Langue Naturelle*, Leuven/Belgique.

KUPSC A. & ABEILLÉ A. (2008). Growing treelex. In A. F. GELBUKH, Ed., *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, p. 28–39 : Springer.

MANGUIN J.-L., FRANÇOIS J., EUFE R., FESENMEIER L., OZOUF C. & SÉNÉCHAL M. (2004). Le dictionnaire électronique des synonymes du crisco : un mode d'emploi à trois niveaux. *Les Cahiers du CRISCO*, **17**.

MESSIANT C., KORHONEN A. & POIBEAU T. (2008). Lexschem : A large subcategorization lexicon for french verbs. In *Language Resources and Evaluation Conference (LREC)*, Marrakech.

PLOUX S. & VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. In *Traitement automatique des langues*, volume 39, p. 161–182.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE E. & BOULLIER P. (2006). The Lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *Proc. of LREC'06*.

SAGOT B. & FISER D. (2008). Building a Free French WordNet from Multilingual Resources. In *Proc. of Ontolex*, Marrakech, Maroc.

VENANT F. (2007). Utiliser des classes de sélection distributionnelle pour désambiguiser les adjectifs. In *Traitement Automatique des Langues Naturelles*, Toulouse, France.

VENANT F. (2008). Représentation géométrique et calcul dynamique du sens lexical : application à la polysémie de livre. *Langages*, **172**.